Senior Capstone Thesis

Gender Bias in Neural Machine Translation

Yuxin Liao

Thesis Advisor: Professor Chris Callison-Burch
Engineering Advisor: Professor Rajeev Alur

University of Pennsylvania
School of Engineering and Applied Science
Department of Computer and Information Science

May 3, 2021

Abstract

With recent advances in deep learning, neural machine translation (NMT) has superseded the statistical approach to machine translation (MT) as the state-of-the-art framework. NMT not only represents a thrilling milestone in MT research, but has also been widely embraced by industrial MT systems and incorporated into tools like Google Translate and Microsoft Translator. Despite its growing popularity and influence, NMT is not free of implicit biases. Since the training corpora for natural language processing (NLP) tasks – including NMT – often reflect societal norms and biases, NMT models are prone to translation inaccuracies that exhibit gender stereotypes. This thesis attempts to shed a light on this issue by offering a summary of recent studies on gender bias in NMT. The first section provides an overview of the NMT architecture and introduces technical concepts such as neural networks, word embeddings, and encoder-decoder models. The second section discusses research that seeks to understand gender bias in the broader context of NLP, while the third section focuses on bias in NMT and approaches to mitigate it.

Table of Contents

Abstract ii

Table of Contents iii

I. Introduction 1

II. Overview of Neural Machine Translation 3

2.1 Neural Networks 3

2.2 Word Embeddings 4

2.3 Encoder-Decoder Models 6

2.3.1 RNN Encoder-Decoder 7

2.4 Attention and Transformers 9

2.4.1 Attention Mechanism 9

2.4.2 Transformer 10

2.5 Performance Evaluation 11

2.5.1 BLEU 11

2.5.2 METEOR 12

2.5.3 TER 13

III. Gender Bias in Pre-trained Models 14

3.1 Gender Bias in Word Embeddings 14

3.2 Evaluation Frameworks 15

3.2.1 Analogy-based Evaluation 15

3.2.2 WinoBias 16

3.2.3 Winogender 18

3.2.4 Word Embedding Association Test (WEAT) 19

3.2.5 Discovery of Correlations (DisCo) 21

3.3 Mitigation Techniques 21

3.3.1 Word Embedding Debiasing 21

3.3.2 Counterfactual Data Augmentation 23

3.3.3 Dropout Regularization 24

IV. Gender Bias in Neural Machine Translation 25

4.1 Challenges in Machine Translation 25

4.2 Evaluation Frameworks 26

4.2.1 WinoMT 26

4.3 Mitigation Techniques 28

4.3.1 Word Embeddings Techniques 29

4.3.2 Domain Adaptation Techniques 30

4.3.3 Cross-lingual Pivoting Technique 31

V. Conclusion 34
5.1 Future Work 34
References 36

I. Introduction

In a world brimming with thousands of languages, human translation has long been our default means to overcome language barriers that hinder cross-cultural communication. However, manual translation is a laborious task. Due to the productive nature of language, translation could be repetitive but never predictable. Furthermore, as language is often intertwined with culture, an adequate translation that captures subtleties and nuances demands not only mastery of source and target languages but also human creativity. Throughout the 17th century, numerous prominent scholars including René Descartes tinkered with the idea of one universal language (Schwartz, 2018). While this effort proved to be futile and arguably impossible, the emergence of machine translation in the mid-20th century provided a promising alternative that might finally be able to tear down linguistic barriers.

Machine translation (MT) can be defined as the use of machines to translate between natural languages (Jurafsky & Martin, 2020, Chapter 11). Prior to the advent of neural machine translation (NMT) around 2014, statistical machine translation (SMT; Brown et al., 1993) had been the dominant framework for over two decades. While SMT relies on massive bilingual corpora to construct a probabilistic model for translation, often in a phrase-by-phrase manner, NMT utilizes recurrent neural networks (RNNs) to encode and decode sequences in an end-to-end fashion (Sutskever et al., 2014; Cho et al., 2014). With improvements over the past years, NMT has surpassed SMT in human evaluations and established a new state of the art (Junczys-Dowmunt et al., 2016; Castilho et al., 2017); in fact, systems like Google NMT (GNMT) have decreased translation errors by an average of 60% on samples in several major language pairs (Wu et al., 2016).

Even though NMT signifies a remarkable milestone in the field of machine translation, like many NLP tasks trained on human language texts (Sun et al., 2019), NMT also exhibits gender bias. For the purpose of this thesis, we define gender bias as the preference or prejudice toward one gender over the other. While often implicit, such bias can manifest in alarming ways, ranging from a tendency to use masculine defaults (Prates et al., 2019) to a reliance on prejudiced (albeit potentially misleading) association between occupation and gender (Rudinger et al., 2018; Zhao et al., 2018a). For example, when translating the English sentence "The doctor asked the nurse to help *her* in the procedure" into Spanish, Google Translate produces "*El* doctor le pidió a *la* enfermera que la ayudara en el procedimiento", identifying the doctor as male in spite of the pronoun "her" (Stanovsky et al., 2019). Inaccuracies like this not only undermine translation quality – especially on a document level – but may also reinforce and even amplify existing societal biases.

This thesis aims to shed a light on this issue by summarizing recent efforts to identify and mitigate gender bias in machine translation, while highlighting limitations and future research directions. In this literature review, we first explore the architecture of NMT and introduce technical concepts such as neural networks, encoder-decoder models, and MT evaluation metrics. We then investigate gender bias in word embeddings and pre-trained language models,

as well as how it propagates to various NLP tasks (Sun et al., 2019), including coreference resolution (Zhao et al., 2019). Using insights from the discussion on gender bias in NLP, we concentrate on gender bias in the domain of machine translation and examine some state-of-the-art studies on this subject matter.

II. Overview of Neural Machine Translation

This section introduces the ideas of neural networks and word embeddings. It explores machine translation techniques such as recurrent neural network (RNN) encoder-decoder, the attention mechanism, and the Transformer architecture. It then discusses some common evaluation metrics of machine translation.

2.1 Neural Networks

As a prelude to the discussion of neural machine translation (NMT) models, we first introduce the concepts of neural networks and word embeddings (Section 2.2), which play a crucial role in many natural language processing tasks including NMT. Neural networks are a family of machine learning algorithms that take in various inputs to predict outputs. Inspired by neurons in the human brain, a neural network consists of computational units that each accept multiple inputs to produce a single output, which can then be passed as input to subsequent units in the network. Compared to linear models that attempt to find a linear relationship between different features, neural networks are non-linear and thus much more versatile. With the use of multiple layers of computing units, neural networks can represent and possibly solve an incredibly large space of problems (Koehn, 2020, Chapter 5).

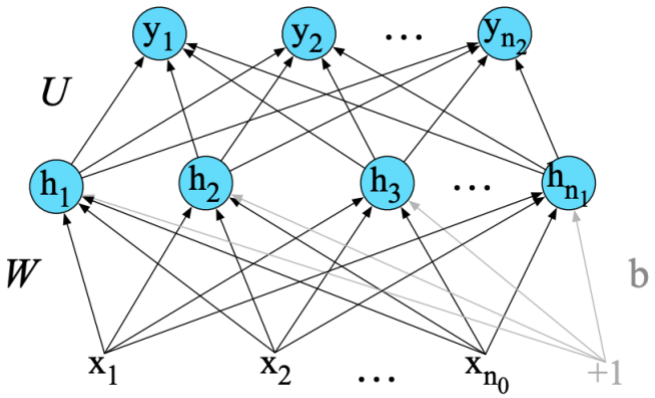

Figure 1 A simple feedforward neural network with one input layer, one hidden layer, one output layer, and one bias term (Jurafsky & Martin, 2020, p.134).

In a standard neural architecture, each layer is fully connected to all nodes from the previous layer and the subsequent layer. The input layer and the hidden layer(s) may also contain bias units, which are helpful in the case where all input values are 0. A feedforward neural network with one hidden layer consists of the following components (Koehn, 2020, p. 68):

- Input vector $\boldsymbol{x} = (x_1, x_2, \ldots x_n)$
- Hidden vector $\boldsymbol{h} = (h_1, h_2, \ldots h_m)$
- Output vector $\boldsymbol{y} = (y_1, y_2, \ldots y_l)$
- Weight matrix $\boldsymbol{W}$ connecting input nodes to hidden nodes

- Weight matrix $\boldsymbol{U}$ connecting hidden nodes to output nodes

For each hidden node, its output can be computed by multiplying the weight matrix $\boldsymbol{W}$ by the input vector $\boldsymbol{x}$, adding a bias term $\boldsymbol{b}$ to the product, and applying an activation function g (e.g., sigmoid, tanh, ReLU) to produce a hidden output $\boldsymbol{h}$, as illustrated in Eq. 1:

$$\boldsymbol{h} = g(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}) \qquad \text{(Equation 1)}$$

For each output node, the weight matrix $\boldsymbol{U}$ is first multiplied by the hidden vector $\boldsymbol{h}$ to produce an intermediate output (Eq. 2). The intermediate output vector is then normalized with the softmax function (Eq. 3) so that the outputs resemble a probability distribution (i.e., all values are between 0 and 1 and sum to 1).

$$\boldsymbol{z} = \boldsymbol{U}\boldsymbol{h} \qquad \text{(Equation 2)}$$
$$\boldsymbol{y} = \text{softmax}(\boldsymbol{z}) \qquad \text{(Equation 3)}$$

For better performance, neural architectures often have multiple hidden layers stacked together, resulting in deeper networks that are utilized in deep learning. In addition to feedforward neural networks, NLP tasks often employ other neural frameworks, such as recurrent neural networks (RNNs; Elman, 1990), Long Short-Term Memory networks (LSTM; Hochreiter & Schmidhuber, 1997), and convolutional neural networks (CNNs; LeCun et al., 1998). Section 2.3 introduces the encoder-decoder architecture for machine translation, with a specific focus on models based on recurrent neural networks.

2.2 Word Embeddings

Word embeddings (and vector semantics in general) are based on the idea that words that are used and occur in similar contexts tend to have similar meanings, formally known as the distributional hypothesis in linguistics (Joos, 1950; Harris, 1954). Leveraging this parallel between distributional and semantic similarities, words can be represented as vectors in a high-dimensional space. For example, semantically related words such as *cat* and *dog* tend to occur in similar environments, so do morphologically related words such as *cat* and *cats*; therefore, these words would have similar vector representations and are close to each other in the vector space (Jurafsky & Martin, 2020, Chapter 6). In addition, once every word in a dictionary is mapped to a numeric vector, we can easily quantify the similarity between words with a distance function such as the cosine similarity metric.

One approach of constructing vectors involves creating co-occurrence matrices (i.e., term-document or term-term matrix) based on the frequency of words in a collection of d documents (Jurafsky & Martin, 2020, Chapter 6). A term-document matrix represents each word as a vector of length d (size of the document collection), while a term-term matrix represents each word as a vector of length |V| (vocabulary size). Weighting algorithms such as tf-idf (term frequency-inverse document frequency) and PPMI (Positive Pointwise Mutual Information) would then be applied to transform counts to word vectors. Since rarely do words co-occur with all other words, co-occurrence matrices often lead to vectors that are long and sparse (i.e., mostly

zeros). To address this, the word2vec software package (Mikolov et al., 2013a; Mikolov et al., 2013b) was introduced to create short dense vectors, also known as embeddings.

Figure 2 Visualization of word embeddings projected into two-dimensional space (Koehn, 2017, p. 36).

Instead of computing frequencies of words, word2vec aims to predict word co-occurrences; one such model is skip-gram with negative sampling (SGNS; Mikolov et al., 2013b). For every word in the text, the algorithm treats the other words in its context window as positive examples of environments in which the target word occurs. To obtain negative samples, for each target-context pair (i.e., positive sample), the algorithm utilizes negative sampling algorithms to randomly select *k* words (i.e., noise words) in the lexicon that do not co-occur with the target. Once positive and negative samples are constructed from a corpus, the algorithm uses logistic regression to train a binary classifier to distinguish between the two cases (Eq. 4 illustrates the loss function for a single target word *w*). Finally, the weights learned during training will be used as word embeddings.

$$L = -[\log P(+|w,c) + \sum_{i=1}^{k} \log(1 - P(-|w,n_i))] \qquad \text{(Equation 4)},$$

where (*w*, *c*) is a single target-context pair from the positive examples, *k* is the number of negative samples for each positive sample, and *n* is one such negative example.

One limitation of word2vec embeddings is that their representation of each word in the vocabulary is static. As a result, they are unable to take a word's syntactic and semantic environment into account, and they often fail to adequately model polysemous words in different linguistic contexts. Contextualized embeddings such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) attempt to address this problem by creating a dynamic representation of

each word. ELMo (Embeddings from Language Models) pre-trains a directional LSTM on a massive text corpus, so that the language model can predict not just the following but also the previous word. Depending on the task, the model learns a function of the hidden states of the entire input sentence, which allows it to create a contextualized representation of each word.

BERT (Bidirectional Encoder Representations from Transformers), on the other hand, uses a stacked Transformer architecture (Section 2.4.2). With the Transformer's self-attention mechanism, the model is able to predict any word based on its surrounding words. Similar to ELMo, BERT is first pre-trained on a large monolingual corpus then fine-tuned for a specific task. Since ELMo and BERT significantly outperform traditional static word embeddings on a variety of NLP tasks, NMT opts for dynamic contextualized embeddings when encoding and decoding words.

2.3 Encoder-Decoder Models

At the sentence-level, machine translation can be defined as the task of converting a sequence of tokens in the source language into a sequence of tokens in the target language. A common architecture for such task is the encoder-decoder network (Cho et al., 2014), which belongs to the broader class of sequence-to-sequence neural network models that map one sequence to another.

While encoder-decoder models may vary based on their underlying language models (for example, various flavors of neural networks such as RNN, LSTM, CNN can be used), a generic encoder-decoder framework has the following three components (Jurafsky & Martin, 2020, pp. 208-209):

1. An encoder that reads in a sequence of tokens of arbitrary length in the source language, $\boldsymbol{x} = (x_1, \ldots, x_n)$, and converts the tokens into a sequence of contextualized representations $\boldsymbol{h} = (h^e_1, \ldots, h^e_n)$.
2. A context vector $\boldsymbol{c}$ that is a function of all the encoder hidden states $(h^e_1, \ldots, h^e_n)$ and captures the essence of the source text.
3. A decoder that reads in the context vector $\boldsymbol{c}$ and produces a variable-length sequence of vector representations, which will then be transformed into tokens in the target language, $\boldsymbol{y} = (y_1, \ldots, y_m)$.

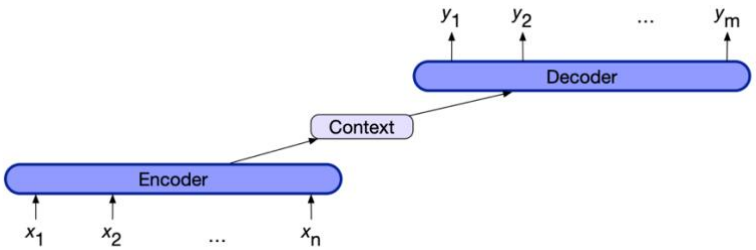

Figure 3 Illustration of an encoder-decoder model. The *context* is a function of the hidden representation of the input and serves as the input to the decoder (Jurafsky & Martin, 2020, p. 208).

To train encoder-decoder models, we use a parallel corpus that contains aligned pairs of sentences in the source and target languages. Each pair of input and output sentences are concatenated with a special token in the middle as a separator. With this training set, the encoder and decoder components of the model are trained jointly to maximize the conditional log-likelihood of a target sentence given a source sentence.

To translate an input sentence, at each step of the decoding process, the algorithm predicts an output word by first computing the probability distribution over all possible outputs (i.e., all words in the vocabulary). The algorithm may take a greedy approach and always select the word with the highest probability at every step. However, this locally optimal choice is rarely optimal for the entire sentence, since the best translation might contain words that initially seem less probable.

One alternative strategy to the greedy method is beam search. Instead of choosing the single best token, beam search keeps track of *k* candidate translations, where *k* is called the beam width. At each timestep, the algorithm selects *k* most likely word choices based on the probability distribution. The algorithm uses distinct decoders to expand all *k* hypotheses to obtain *k* * |V| candidates, where |V| is the size of the vocabulary. Each of these candidates will be scored with $P(y_i|x, y_{<i})$, which evaluates the probability of a potential output given the path that led to it. The search space is then pruned down to *k* best hypotheses based on this score, so that the number of hypotheses on the search frontier does not exceed the beam width.

The decoding process proceeds until an end-of-sentence token is generated for each of the *k* hypotheses. A score is then computed for a hypothesis *y* with Eq. 5. Note that the score is normalized by dividing the negative log probability by the length of *y* to compensate for the fact that longer strings tend to receive lower probabilities. Depending on the use case, either the best translation or a subset of the top translations will be returned.

$$score(y) = -\log P(y|x) = \sum_{i=1}^{t} \log P(y_i|y_1, \ldots, y_{i-1}, x) \qquad \text{(Equation 5)}$$

2.3.1 RNN Encoder-Decoder

A popular underlying model for encoder and decoder is recurrent neural networks (RNNs), which are a type of neural networks that incorporates the previous hidden state h_{t-1} when computing hidden state h_t at time step *t*. Given the complex syntactic and semantic structures that exist in natural languages, RNNs often stand out as a suitable choice. Since information obtained in previous timesteps is continuously embedded into the current hidden state, RNNs can not only accept input sentences of an arbitrary length but also capture the long-distance dependencies between words. This is crucial for NMT since a faith translation must preserve agreement in gender, number, and other grammatical properties, regardless of the number of intervening words in a sentence.

An RNN-based encoder-decoder framework is usually composed a pair of RNNs. Fig. 4 is an illustration of a model with a single-layered RNN as the encoder and the decoder.

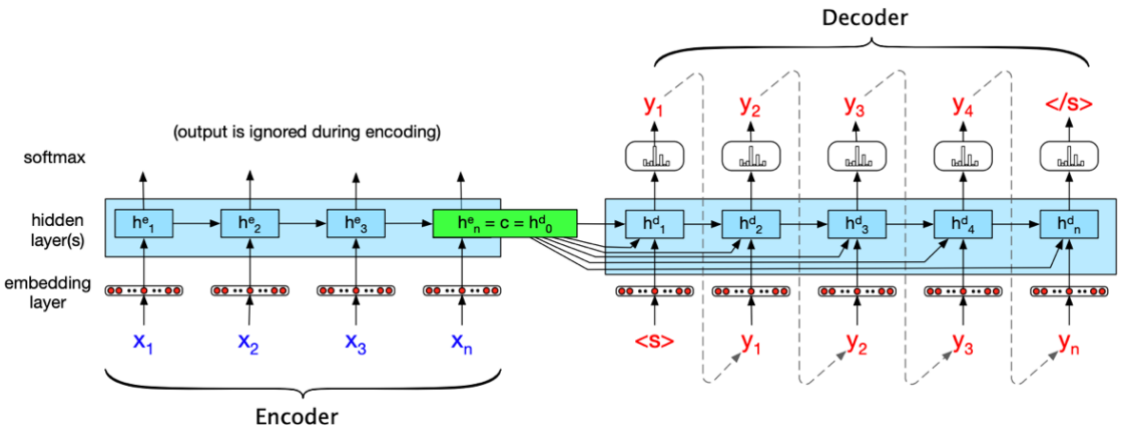

Figure 4 Illustration of a singled-layered RNN encoder-decoder model (Jurafsky & Martin, 2020, p.210).

Given an input sequence in the source language $\boldsymbol{x} = (x_1, x_2, x_3, \ldots, x_n)$, where each x_i is the embedding of a word in the sentence, the encoder RNN maintains a hidden state vector, $\boldsymbol{h^e}$, that is continuously updated as the encoder reads over $\boldsymbol{x}$ sequentially. At each timestep t, the RNN updates $\boldsymbol{h^e}$ according to the following equation:

$$h_t^e = f(h_{t-1}^e, x_t) \qquad \text{(Equation 6)},$$

where f is a non-linear activation function such as a sigmoid function or long short-term memory (LSTM), h_{t-1}^e is the cumulative hidden state from the prior state, and x_t is the word embedding of the input token at the current timestep t.

At the end of this encoding process (indicated by the end-of-sentence symbol), the latest hidden state of the encoder would contain contextualized information of the entire input sentence derived from all previous timesteps. This hidden state is often designated as the context vector $\boldsymbol{c}$ that will then be passed onto the decoder.

As part of the decoding phase, we initialize the decoder RNN's hidden state h_0^d with the context vector $\boldsymbol{c}$ generated by the encoder. At each timestep t, the decoder uses Eq. 7 to update the hidden state h_t^d and Eq. 8 to compute a probability that will then be used to generate an output token, until the end-of-sentence token is generated. Note that to prevent the influence of $\boldsymbol{c}$ from being diluted as more output is generated, $\boldsymbol{c}$ is available as a parameter to f and g throughout the decoding process.

$$h_t^d = f(h_{t-1}^d, y_{t-1}, \boldsymbol{c}) \qquad \text{(Equation 7)},$$

where f is an activation function.

$$\mathrm{P}(\mathrm{y_t}|y_{t-1}, y_{t-2}, \ldots, y_1, \boldsymbol{c}) = g(h_t^d, y_{t-1}, \boldsymbol{c}) \qquad \text{(Equation 8)},$$

where g is an activation function that produces valid probability distribution, such as a softmax function.

2.4 Attention and Transformers

2.4.1 Attention Mechanism

One of the challenges of machine translation is capturing the long-distance dependencies that exist in a sentence. The standard encoder-decoder architecture attempts to address this problem by compressing an input sequence of arbitrary length into a context vector of fixed length, so that the context vector embodies the contextualized information of the entire sentence. However, the performance of such models tends to deteriorate rapidly as the length of the input sentence increases. This effect can be attributed to the limited representation power of a fixed-length intermediate vector: since the standard model uses only the final hidden state of the encoder (or some function of the hidden states) as the context vector, the beginning of the input may not be equally well represented, and relevant parts of the input may not be emphasized enough when generating its corresponding translation.

The attention mechanism – first proposed by Bahdanau et al. in 2014 – presents a potential solution to this problem. Instead of representing the entire input sequence with a static, fixed-length vector, the context vector for each time step of the decoding phase can be computed dynamically. By focusing on parts of the input that are relevant for a particular output token, the attention mechanism can better capture long-distance dependencies for each word in the output.

A standard attention-based model first creates a set of hidden states to represent the input by running a bidirectional RNN (BiRNN) on the source sentence. At each time step *t* in the encoding process, we concatenate the two hidden vectors (one from each direction) to create a bidirectional representation of the token's context. We then concatenate all the hidden states generated at each time step into a matrix $\boldsymbol{H^e}$ where each column corresponds to the hidden state of an input token. Now instead of a single fixed context vector, the decoder can reference this matrix representation of the input when deriving the context vector at each time step of the decoding phase.

To generate context matrix c_t at each time step *t*, the decoder must compute an attention vector α_t that captures the relevance of each encoder hidden state such that $c_t = \boldsymbol{H^e}\alpha_t$. This process consists of the following steps:

1. For each encoder hidden state, h_i^e, compare the similarity between h_i^e and the prior decoder hidden state h_{t-1}^d by computing an attention score $score(h_{t-1}^d, h_i^e)$. The scoring function can simply be the dot product between h_i^e and h_{t-1}^d, as long as the two vectors have the same length. We can also use a more sophisticated scoring function such as the bilinear function, which adds a learnable parameter W_s to the equation. The weight vector not only increases the expressiveness of the scoring function, but also allows vectors to have different dimensions (unlike the dot product).

2. After computing a score for each vector in $\boldsymbol{H^e}$, we can normalize the scores into a probability distribution with softmax to obtain the proportional relevance of h_i^e to h_t^d:

$$\alpha_{it} = \text{softmax}\left(score\left(h_{t-1}^{d}, h_i^e\right) \forall\, i \in e\right) = \frac{\exp(score\left(h_{t-1}^{d}, h_i^e\right))}{\sum_k \exp(score\left(h_{t-1}^{d}, h_k^e\right))} \qquad \text{(Equation 9)}$$

3. Compute the weighted average of all the encoder hidden states, which will then be used as the context vector for the current time step *t*.

$$c_t = \sum_i a_{it} h_i^e \qquad \text{(Equation 10)}$$

2.4.2 Transformer

One drawback of an RNN-based encoder-decoder model is that recurrent neural networks require a sequential walk-through of the entire input in order to generate hidden states. This approach is not only time and space consuming but might also lead to a loss of relevant information over a long sequence of recurrent connections. The Transformer (Vaswani et al., 2017), on the other hand, is a non-recurrent and highly parallelized architecture that solely relies on the attention mechanism. It's not only more time and space efficient but also established a new state-of-the-art BLEU score when it was first introduced in 2017.

In addition to components such as feedforward neural networks and the encoder-decoder attention mechanism (Section 2.4.1), the Transformer also utilizes self-attention in its encoder and decoder. Instead of a recurrent approach, self-attention creates a representation of a sequence by computing the relationship between a target position and other positions in the same sequence. By attending to relevant contextual information for each input token, the model can better identify its long-distance dependencies and provide a more refined representation of the word. Moreover, since the computation for each position is independent of other computations, the self-attention mechanism can perform information extraction and inference for large contexts by utilizing parallel computation.

To facilitate learning during training, the embedding of each input word is associated with three weight matrices: Query vector, Key vector, and Value vector. Eq. 11 can be used to compute the outputs of a self-attention layer. Furthermore, to capture both the local context and long-range dependencies of a word, multi-head self-attention – which consists of multiple self-attention layers running in parallel – may be used to extract different kinds of relations. These parallel layers each have a set of distinct parameters but are focused on learning different aspects of the context. The output from each layer can be concatenated and reduced to the original output dimension, as expected by the subsequent layer.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad \text{(Equation 11),}$$

where d_k is the dimension of the query and key vectors.

$$\text{MultiHead}(Q, K, V) = \boldsymbol{W^o}(head_1 \oplus head_2 \ldots \oplus head_h) \qquad \text{(Equation 12),}$$

where $head_i$ is the output of i^{th} head (computed with Eq. 11) and $\boldsymbol{W^o}$ is the weight matrix that projects the concatenated output to the original output dimension.

The output of the self-attention layers will then be passed to additional feedforward layers, residual connections, normalization layers. These layers form a transformer block that can be stacked on top of each other. Furthermore, since the order of tokens is no longer preserved, the Transformer uses positional encoding to combine information about the relative positions of words with their embeddings, and such information will be passed to the transformer blocks.

As in RNN encoder-decoder models, the Transformer uses attention between its encoder and decoder to focus on relevant parts of the input sequence. The decoder also utilizes stacked self-attention layers, in addition to residual connections, layer normalization, and a feedforward network.

2.5 Performance Evaluation

To track progress in the field machine translation, it is crucial to develop systematic ways to evaluate the effectiveness of different translation systems. Currently, human evaluation sets the gold standard for translation assessment. One human assessment approach involves grading outputs along two dimensions: fluency and adequacy (Jurafsky & Martin, 2020, p. 221). While fluency is concerned with how natural and grammatically correct the translation is, adequacy is concerned with how well the translation captures the meaning of the input sentence. Alternatively, human annotators can rank outputs from different systems on a sentence-by-sentence basis to determine the superior translation. An obvious downside of human assessment is that it could be inefficient and expensive, which means automatic evaluation methods are much more practical and preferable, especially when it comes to evaluating large amounts of translations across different system configurations. This section explores a few popular automatic metrics, including BLEU, METEOR, and translation error rate (TER).

2.5.1 BLEU

Developed by IBM researchers Papineni et al. in 2002, BLEU (Bilingual Evaluation Understudy) is the most widely used automatic evaluation metric. Given a candidate MT translation and a gold standard produced by manual translation, BLEU computes the number of n-gram (word-sequence) overlaps between the two sentences.

$$prec_n = \frac{\sum_{S \in C} \sum_{\text{n-gram} \in S} \text{count\_match(n\_gram)}}{\sum_{S' \in C} \sum_{\text{n-gram}' \in S'} \text{count\_all(n\_gram}')} \quad \text{(Equation 13),}$$

where C is the set of all candidate sentences in the corpus; S is a candidate sentence for which we count all the n-gram matches of a certain type (e.g., unigram matches) with the reference. The denominator is simply the total of all such n-grams (e.g., unigrams) in all candidate sentences.

Since this metric is precision-focused, it is possible to game the system by producing very short outputs consisting only of n-grams with high confidence (Dorr et al., 2010). Instead of

using recall (Eq. 15) or F-measure (Eq. 16), BLEU uses a brevity penalty (Eq. 17) to explicitly address such scenarios, and the BLEU metric is defined in Eq. 18.

$$Precision = \frac{\text{matches}}{\text{n-gram\_count}(c)} \qquad \text{(Equation 14)},$$

where the numerator is the number of n-gram matches between the candidate and the reference, and the denominator is the total number of n-grams in the *candidate* sentence.

$$Recall = \frac{\text{matches}}{\text{n-gram\_count}(r)} \qquad \text{(Equation 15)},$$

where the numerator is the number of n-gram matches between the candidate and the reference, and the denominator is the total number of n-grams in the *reference* sentence.

$$F_1 = \frac{2PR}{P+R} \qquad \text{(Equation 16)},$$

where P denotes precision (Eq. 14) and R denotes recall (Eq. 15).

$$BP = \min(1, \exp\left(1 - \frac{r}{c}\right)) \qquad \text{(Equation 17)},$$

where r is the sum of the length of all references, and c is the sum of the length of all candidates.

$$BLEU = BP \times (\prod_{n=1}^{4} prec_n) \qquad \text{(Equation 18)},$$

where n specifies the n-gram length[1].

2.5.2 METEOR

Unlike BLEU, which only considers exact matches between the translation and its reference, METEOR (Banerjee & Lavie, 2005) is more focused on the semantic quality of a sentence and allows more variations in word choice. In addition to exact matching, METEOR also uses stem matching and synonym matching to detect morphologically or semantically similar words, and synonyms can often be obtained via WordNet (Fellbaum, 1998).

Furthermore, while BLEU relies solely on a fixed brevity penalty to compensate for the lack of recall, METEOR incorporates both precision and recall into the equation to compute their harmonic mean:

$$F_{mean} = \frac{P \cdot R}{\alpha P + (1-\alpha) R} \qquad \text{(Equation 19)},$$

where P denotes precision (Eq. 14), R denotes recall (Eq. 15) and α is usually set to be 0.9 since METEOR is a recall-focused metric.

[1] A common maximum length for n-grams is four.

Finally, METEOR also uses a penalty score to favor longer matches over shorter fragments (e.g., unigrams):

$$penalty = \gamma \cdot frag^{\beta} \qquad \text{(Equation 20)},$$

where $frag$ is the number of consecutive unigram matches divided by the total number of unigram matches; γ is defaulted to be 3.0 and β is 0.5.

The METEOR score can be defined as:

$$score = (1 - penalty) \cdot F_{mean} \qquad \text{(Equation 21)}$$

2.5.3 TER

Translation error rate (TER; Snover et al., 2006) is concerned with the number of edits, or movement or words, from the candidate translation to reference. An edit could be an operation that inserts, deletes, or substitutes a word, or an operation that shifts a sequence of words. TER is then computed as the total number of substitutions, insertions, deletions, and shifts, over the average number of words across multiple reference sentences.

$$TER = \frac{\text{minimum \# of edits}}{\text{average \# of reference words}} \qquad \text{(Equation 22)}$$

It's worth noting that the computation of TER score is NP-complete, since there are too many shifts to efficiently compute the optimal number edits. As a result, a greedy algorithm is often used to first compute the word error rate (WER) between a translation and its reference, and then iteratively find the best shift operation until the score decreases.

III. Gender Bias in Pre-trained Models

This section explores gender bias in pre-trained language models, with a focus on bias in word embeddings. Since word embeddings are widely used in various NLP tasks, understanding how human biases are encoded into these systems might shed a light on bias in machine translation. Furthermore, this section discusses methods for quantifying bias in pre-trained models, as well as techniques that can be employed to mitigate gender bias.

3.1 Gender Bias in Word Embeddings

Word embeddings play a crucial role in a variety of natural language processing applications, and the development of word2vec and contextualized embeddings (e.g., ELMo, BERT) led to significant improvements in tasks including machine translation. However, as pre-trained models are derived from corpora of human language texts, word embeddings have been shown (Bolukbasi et al., 2016; Zhao et al., 2019; Webster et al., 2020) to exhibit biases that reflect stereotypes about gender, race, age, and other demographic factors.

In 2016, Bolukbasi et al. published a seminal paper that revealed human-like biases in w2vNEWS embedding – a 300-dimensional word2vec embedding trained on a Google News corpus of 3 million English words. Since vectors are used to embody word semantics, the relationship between words can be modeled by the difference between corresponding word vectors. For instance, when solving the analogy puzzle "man is to king as woman is to *x*" (denoted as *man:king :: woman:x*), word embedding arithmetic returns *x = queen*, since $\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$

Even though vector arithmetic can capture useful relationships between words, it can also shed a light on biases encoded into the system. By projecting word2vec embeddings of occupations onto the she-he axis, the model generates analogies such as *man:computer scientist :: woman:homemaker* and *father:doctor :: mother:nurse*. In a similar fashion, embeddings can make biased predictions based on names and are likely to rank "computer science" as more highly related to male names than female names. This correlation between occupation and gender word embeddings might result in amplified bias in downstream applications. For instance, if word vectors were employed to improve the relevance of search results (Nalisnick et al., 2016), the search engine might rank pages related to male computer scientists higher than those related to female computer scientists. Furthermore, comparable results were replicated by the GloVe word embeddings pretrained on a web-crawl corpus (Bolukbasi et al., 2016), suggesting that such bias is prevalent across different word embedding systems.

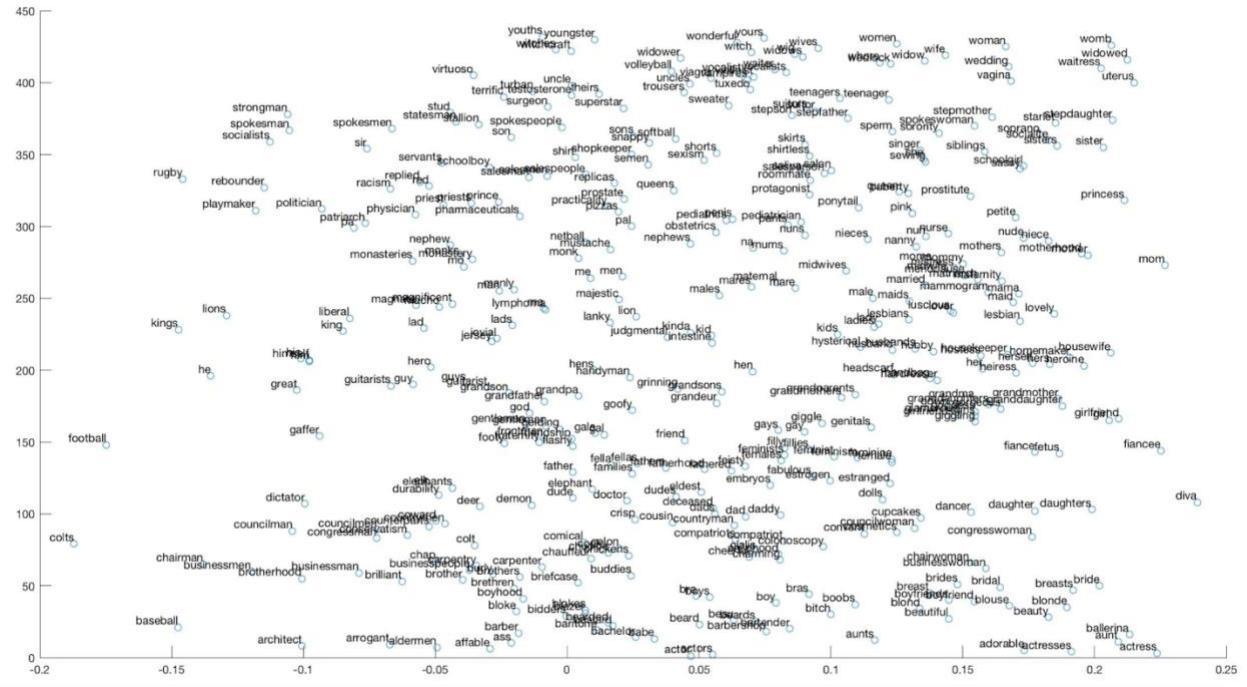

Figure 5 Projection of potentially stereotyped words (represented by GloVe vectors) along the she-he axis. Words to the left are he words and words on the right are she words (Chakraborty et al., 2016).

While contextualized embedding models such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2019) and ALBERT (Lan et al., 2020) have led to substantial improvement across a range of NLP tasks, they are not free from gender bias. Zhao et al. (2019) show that not only does the training corpus for ELMo (i.e., the Billion Word corpus) contain three times more masculine pronouns (i.e., *he, his,* and *him*) than feminine pronouns (i.e., *she, her,* and *hers*), masculine pronouns are also more likely to co-occur with occupation words. In addition to the bias in training data, ELMo seems to represent and propagate information associated with male entities better than information associated with female entities. In their study, Zhao et al. show that an SVM classifier is 14% more accurate when predicting the gender of a profession term in a sentence (e.g., "the engineer went back to her home") when the true gender is male instead of female. BERT and ALBERT also learn gendered correlations and exhibit quantifiable bias (Webster et al., 2020), which might manifest in downstream tasks such as gender pronoun resolution and coreference resolution.

3.2 Evaluation Frameworks

3.2.1 Analogy-based Evaluation

To detect gender bias in word2vec embeddings, Bolukbasi et al. (2016) propose an analogy-based evaluation approach. They first distinguish between gender-specific words, which are associated with a gender by definition (e.g., *brother*, *sister*), and gender-neutral words, which

encompass all other words and are the target of the debiasing algorithm (described in Section 3.3.1).

Bolukbasi et al. then identify two types of bias: direct bias refers to the association between a gender-neutral word and a clear gender pair (e.g., *he* and *she*), which can be reflected by the locations of vectors (e.g., *doctor* is closer to *man* than *woman*). A less obvious kind of bias, indirect bias, is manifested in the association between gender-neutral words, which is most likely derived from the words' respective relationships with gendered words. For instance, the fact that *receptionist* is closer to *softball* than *football* may be attributed to their respective associations with gender specific words such as *she* and *woman*.

To demonstrate that word embeddings contain quantifiable gender bias in their geometry, Bolukbasi et al. utilize a gender subspace, which consists of the difference vectors of multiple gender pairs (e.g., *she* and *he*, *her* and *his*, *woman* and *man*). Since the gender subspace encompasses multiple directions in which word embeddings could carry gender bias, it is expected to capture the overall gender bias in the embeddings.

Given a set of gender-neutral words, denoted by *N*, and a direction for the gender subspace, denoted by *g*, direct bias can be defined by the following equations:

$$\text{DirectBias}_c = \frac{1}{|N|}\sum_{w \in N} |\cos(w, g)|^c \qquad \text{(Equation 23)}$$

where *c* is a parameter for specifying how strict the bias measurement should be, with $c = 0$ being the strictest (i.e., no overlap between *w* and *g* is tolerated) and $c = 1$ being more lenient. In Bolukbasi et al.'s study with 327 occupation words, $\text{DirectBias}_1 = 0.08$, suggesting that the embeddings of many occupation words have a component along the gender direction.

To measure indirect bias, a more sophisticated equation is used to detect the effect of g on pairs of gender-neutral words (denoted w, v):

$$\beta(w, v) = \left(w \cdot v - \frac{w_\perp \cdot v_\perp}{||w_\perp||_2 ||v_\perp||_2}\right) / \, w \cdot v \qquad \text{(Equation 24)}$$

in which $w_\perp = w - w_g$ and $w_g = (wg)g$. The numerator represents the change in inner product after gender subspace is projected out from w and v, and the overall metric is the ratio of this difference to the original inner product value.

Using this evaluation metric, the indirect bias β between *softball* and *receptionist*, *waitress*, and *homemaker* are 67%, 35%, 38%, respectively, suggesting that g has a substantial impact on the relationship between two supposedly gender-neutral words.

3.2.2 WinoBias

WinoBias (Zhao et al., 2018a) is a challenge dataset for analyzing gender bias in coreference resolution, which is the task of identifying all phrases (i.e., mentions) that refer to the same entity. The WinoBias corpus contains 3160 Winograd-style (Rahman & Ng, 2012) sentences, in which a pronoun must correspond to one of two previously mentioned entities. All sentences

contain references to one of 40 occupations that are considered stereotypical based on US Department of Labor statistics[2] and contain roughly follow two prototypical templates (Zhao et al., 2018a):

Type 1 (WinoBias-knowledge): [entity1] [interacts with] [entity2] [conjunction] [pronoun] [circumstances].
Coreference resolution requires world knowledge about given circumstances since sentences of this type contain no syntactic cues.

Type 2 (WinoBias-syntax): [entity1] [interacts with] [entity2] and then [interacts with] [pronoun] for [circumstances].
Coreference decisions can be made based on syntactic information and understanding of the pronoun, and both semantic and syntactic cues are able to assist with disambiguation. Coreference resolution systems are expected to do well in these cases.

Each sentence in the challenge set can be characterized as either *pro-stereotypical*, in which the pronoun refers to an occupation dominated by the gender of the pronoun, or *anti-stereotypical*, in which the occupation is not dominated by the gender of the pronoun. A coreference resolution system is considered gender biased if it links pronouns to occupations more accurately in pro-stereotypical examples than in anti-stereotypical examples.

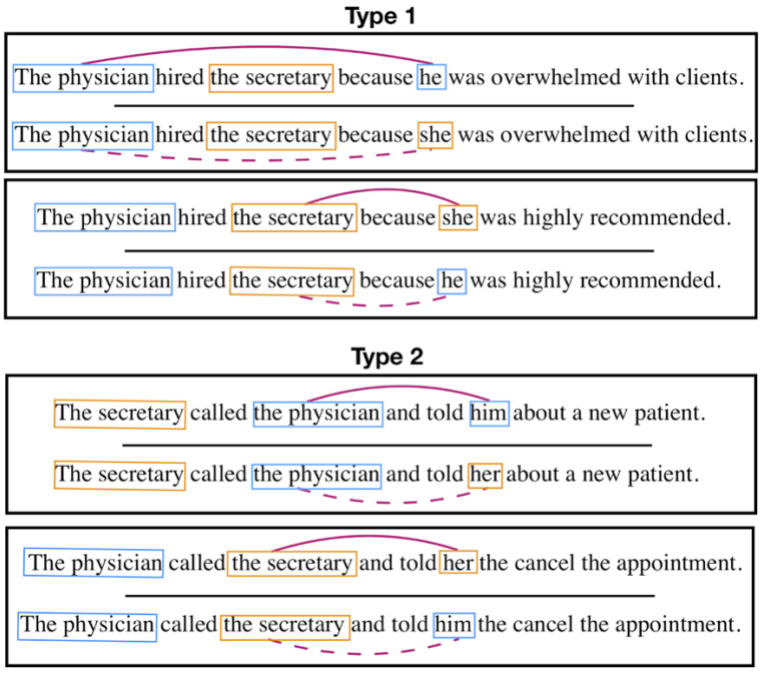

Figure 6 Pairs of gender balanced co-reference tests in the WinoBias dataset. Male and female entities are marked in blue and orange, respectively. For each example, the gender of the pronominal reference is irrelevant for the co-reference decision. Systems must be able to make correct linking predictions in pro-stereotypical scenarios (solid purple lines) and anti-stereotypical scenarios (dashed purple lines) equally well to pass the test (Zhao et al., 2018a).

[2] https://www.bls.gov/cps/cpsaat11.htm

With the WinoBias challenge set, the study reveals that three systems representative of three paradigms – the Stanford Deterministic Coreference System (rule-based; Raghunathan et al., 2010), the Berkeley Coreference Resolution System (feature-driven; Durrett & Klein, 2013), and the UW End-to-end Neural Coreference Resolution System (neural system; Lee et al., 2017) – all contain gender bias. More specifically, all three systems exhibit a significant ($p < .05$) difference in F_1 score between pro-stereotypical and anti-stereotypical conditions, with an average difference of 21.1 across the systems.

Since the training corpus used by these systems, Ontonotes 5.0[3], contains significantly more male entities than female entities, Zhao et al. employ a data augmentation approach with anonymization (discussed in Section 3.3.1) to eliminate this gender imbalance in training data. Furthermore, since coreference systems are built on word embeddings that have been shown to display gender bias (Bolukbasi et al., 2016), Zhao et al. replace GloVe embeddings with debiased vectors (described in Section 3.2.1) to prevent bias propagation. Overall, the combination of these debiasing methods result in significant reduction in bias when evaluated on WinoBias, without significantly affecting accuracy.

3.2.3 Winogender

Similar to WinoBias, Winogender[4] (Rudinger et al., 2018) is a Winograd schema-style challenge set for detecting gender bias in coreference resolution systems. However, the two datasets differ in that Winogender schemas also include gender-neutral pronouns and are human-validated on Mechanical Turk. In addition, each Winogender sentence contains only one occupational mention, as opposed to two.

To reveal cases in which a system may be more or less likely to correctly identify the reference of a pronoun, Winogender utilizes 60 one-word occupations and 120 hand-written sentence templates. Each sentence consists of an occupation, a secondary participant, and a pronoun (i.e., which could be female, male, or gender-neutral) that is coreferential with either the occupation or the participant.

(1a) **The paramedic** performed CPR on the passenger even though she/he/they knew it was too late.

(2a) The paramedic performed CPR on **the passenger** even though she/he/they was/were already dead.

(1b) **The paramedic** performed CPR on someone even though she/he/they knew it was too late.

(2b) The paramedic performed CPR on **someone** even though she/he/they was/were already dead.

Figure 7 Example sentences generated from the occupation paramedic. Correct answers are highlighted in bold (Rudinger et al., 2018).

[3] https://catalog.ldc.upenn.edu/LDC2013T19
[4] https://github.com/rudinger/ winogender-schemas

With this dataset, Rudinger et al. find that across all three types of coreference system architectures (rule-based, statistical, and neural), male pronouns are more likely to be resolved as the occupation than female or neutral pronouns; meanwhile neutral pronouns (i.e., *they/their/them*) are often failed to be resolved correctly, possibly due to ambiguity in number agreement. Furthermore, as illustrated in Figure 8, the gender preferences exhibited by all three systems mirror real-world occupational gender disparities, based on both the U.S. Bureau of Labor Statistics and Bergsma and Lin (2016)'s work on assigning a count-based gender score to a large list of English nouns.

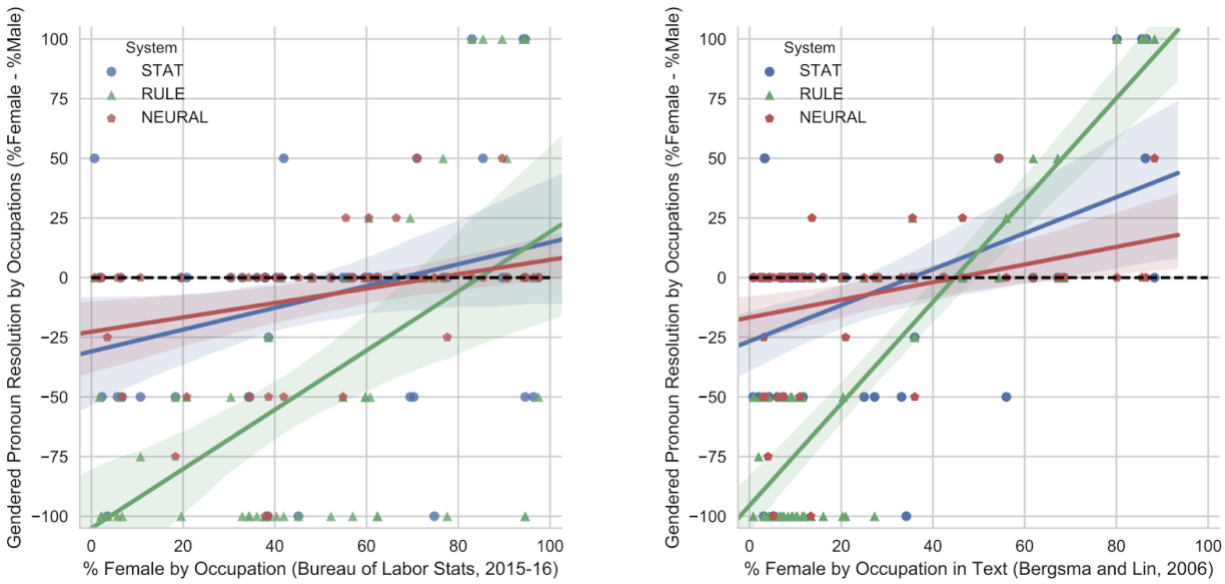

Figure 8 Each point in the plot represents an occupation. The y-axis represents how much more a system prefers a female pronoun to a male pronoun, with 100% being the maximum female bias, and -100% being the maximum male bias. Since the Winogender dataset is gender-balanced for each occupation, the dotted black line at y=0 represents the ideal system with 100% accuracy (Rudinger et al., 2018).

3.2.4 Word Embedding Association Test (WEAT)

The Word Embedding Association Test (WEAT) is a statistical method devised by Caliskan et al. (2017) to quantify social bias in word embeddings. Inspired by the Implicit Association Test[5] (IAT; Greenwald et al., 1998) – a widely used test for measuring human biases, WEAT compares two sets of target concepts of equal size, denoted *X* and *Y*, with two sets of attribute words, denoted *A* and *B*. For instance, *X* and *Y* could be sets of gendered terms such as {*man, male*} and {*woman, female*}, while *A* and *B* are career or family-related words such as {*programmer, engineer, scientist*} and {*nurse, teacher, librarian*}. To assess whether there exists a significant difference between how the two sets of target words relate to the two sets of attribute words, Caliskan et al. use cosine similarity between target and attribute word

[5] https://implicit.harvard.edu/implicit/takeatest.html

embeddings to quantify bias. More specifically, Eq. 25 and Eq. 26 are used to compute the test statistics, Eq. 27 is used to compute the *p*-value, and Eq. 28 is used to compute the effect size.

$$s(t, A, B) = [\text{mean}_{a \in A} \text{sim}(t, a) - \text{mean}_{b \in B} \text{sim}(t, b)] \qquad \text{(Equation 25)},$$

where sim is the cosine similarity between the embeddings for a target word and an attribute word.

$$S(X, Y, A, B) = [\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)] \qquad \text{(Equation 26)},$$

where the degree of bias for each target word in X and Y (i.e., $s(x, A, B)$ and $s(y, A, B)$) is computed with Eq. 25.

$$p = \Pr[S(X_i, Y_i, A, B) > S(X, Y, A, B)] \qquad \text{(Equation 27)}$$

$$d = \frac{S(X,Y,A,B)}{\text{std}_{t \in X \cup Y}\, s(t,A,B)} \qquad \text{(Equation 28)}$$

While the cosine-based methods may expose bias in traditional word embeddings such as word2vec and GloVe, they are less effective when detecting bias in contextualized embeddings such as BERT (May et al., 2019). To quantify bias in BERT, Kurita et al. (2019) propose an alternative algorithm to compute the association between a target and an attribute. The following procedure is illustrated with examples in which [TARGET] = *male gender* and [ATTRIBUTE] = *programmer*.

1. Using the target and the attribute, create a template sentence such as "[TARGET] is a [ATTRIBUTE]" (e.g., "he is a programmer").
2. Replace [TARGET] with masked tokens, denoted by [MASK], to obtain "[MASK] is a [ATTRIBUTE]" (e.g., "[MASK] is a programmer"), denote this as *sentence-1*.
   Query BERT with *sentence-1* and compute p_{target} = P([MASK] = [TARGET] | *sentence-1*), which represents the association between the target and the attribute.
3. Replace both [TARGET] and [ATTRIBUTE] with [MASK] (e.g., "[MASK] is a [MASK]"), denote this as *sentence-2*.
   Query BERT with *sentence-2* and compute prior probability p_{prior} = P([MASK] = [TARGET] | *sentence-2*), which represents how likely the target word is in BERT given the sentence structure alone.
4. Compute the log probability bias score = $\log(\frac{p_{target}}{p_{prior}})$, which indicates the relative association between the target and the attribute (e.g., how much *more* BERT prefers the association between *male gender* and *programmer* than the association between *female gender* and *programmer*).

While conducting WEAT on BERT fails to find any statistically significant biases at $p < 0.01$, the algorithm for log probability bias score queries the underlying language model with masked sentences, and it reveals that BERT also exhibits human-like biases. Furthermore, Kurita et al. show that the gender bias encoded in BERT might influence downstream tasks such as Gendered Pronoun Resolution (GPR; Webster et al., 2018), which is a subtask in coreference resolution. By employing the log probability bias score method, they observe that it may be challenging for BERT to perform coreference resolution correctly when the pronoun is female and the topic is biased towards the male gender.

3.2.5 Discovery of Correlations (DisCo)

With a focus on intrinsic gendered correlations, Webster et al. (2020) propose a new evaluation framework called Discovery of Correlations (DisCo) that reveals and quantifies correlates of gender in contextual representations. DisCo utilizes a series of templates with unfilled slots (e.g., "[PERSON] studies [BLANK] at college"). In each sentence, the [PERSON] slot is filled manually, with either a name (e.g., *Maria*) or a term (e.g., *The poetess*) that is associated with a gender. While the two sources – names from the US Social Security name statistics[6] (denoted Names) and terms from the list of gendered nouns compiled by Zhao et al. (2017; denoted Terms) – only have binary labels (*female* and *male*), the word lists could be extended to include gender-neutral values. The second slot, labeled [BLANK], is then filled by a pre-trained model.

Based on the value of [PERSON], the model might exhibit preference for one gender over another when supplying a word for [BLANK]. Therefore, the DisCo metric is defined to be the number of fills that are significantly associated with a gender over the total number of templates, where the presence of gendered correlations is determined by a χ-squared test.

Unlike evaluation methods based on tasks such as coreference resolution, DisCo provides insight into gendered correlations intrinsic to a pre-trained language model. In the experiments conducted by Webster et al., a large BERT model (with 334 parameters) has a DisCo value of 1.0 when evaluated with Terms and a value of 3.4 when evaluated with Names. Since experiments with random groups (not categorized by gender) achieved a DisCo value of either 0.0 or 0.1, the non-zero DisCo values of BERT suggest that gendered correlations are inherent in the model.

3.3 Mitigation Techniques

3.3.1 Word Embedding Debiasing

To address gender bias in word2vec word embeddings, Bolukbasi et al. (2016) propose a linear algebraic debiasing technique to mitigate bias. The debiasing algorithm has two flavors. In the hard debiasing (also known as "neutralize and equalize") approach, all gender-neutral words are neutralized so that they are all at position **0** in the gender subspace. The equalization step then

[6] https://www.ssa.gov/oact/babynames/limits.html

ensures that gender-neutral words are equidistant from words in gender-specific pairs. For instance, given two gender specific pairs {*grandmother, grandfather*} and {*guy, gal*} and a word gender-neutral word *babysit*, equalization guarantees that *babysit* is equidistant from the words in each pair, though it is plausible that *babysit* is closer to {*grandmother, grandfather*} than {*guy, gal*}. However, one disadvantage of this hard debiasing method is that by completely subtracting gender associations from all gender-neutral words, it might also remove useful semantic differences between words in the same gender specific pair. For example, *grandfather* carries a gender-neutral meaning (i.e., "to permit to continue under a grandfather clause"[7]) that cannot be substituted by *grandmother*, and this distinction is removed during equalization.

In the soft-debiasing approach, the algorithm not only projects embeddings into a subspace orthogonal to the gender subspace, but it also strives to maintain as much similarity to the original embeddings as possible. A parameter is trained to find a balance between preserving the pairwise inner products of embeddings and minimizing the gendered component in gender-neutral words.

To assess the effectiveness of these debiasing algorithms, Bolukbasi et al. use the analogy generation task as described in Section 3.2.1. While initially 19% of the top 150 analogies were perceived to display gender stereotypes by human evaluators, after applying the hard debiasing technique, only 6% of the analogies generated by the new word embeddings were judged as stereotypical. For instance, given the puzzle *man:doctor :: woman:x*, the original embeddings return *x = nurse,* while the debiased embeddings yield *x = physician*. At the same time, the new embeddings are attuned to gender appropriate analogies, as suggested by examples such as *he:prostate cancer :: she:ovarian cancer*. Even though it is more challenging to quantify the impact of debiasing on indirect bias, some qualitative improvements are observed. However, the study also finds that soft debiasing is less effective in reducing gender bias in word embeddings than hard debiasing.

In spite of the effectiveness of strict debiasing, there are some limitations with this approach. First, this method hinges on the correct classification of gender-neutral and gender-specific words; errors in this step might affect the rest of the pipeline and the entire model. Second, as mentioned above, hard debiasing completely removes all gender information from words, including that might be useful or even essential in domains such as medicine and social science (Zhao et al., 2018b). Furthermore, Gonen and Goldberg (2019) argue that simply projecting out a gender direction is rather superficial and does not address the underlying source of bias. While previously "biased" words have an altered geometry with respect to the gender subspace, the relative spatial relationship between word that exhibit a specific bias stays largely the same; as a result, there still might be lingering hidden bias among debiased words. Last, both the analogy-based evaluation framework and the hard debiasing approach cannot be effectively generalized to contextualized word embeddings (Webster et al., 2020), which have become increasingly more popular and widely adopted.

[7] https://www.merriam-webster.com/dictionary/grandfather

3.3.2 Counterfactual Data Augmentation

An alternative to word embeddings debiasing is counterfactual data augmentation (CDA; Lu et al., 2018), which aims to minimize associations between gendered and gender-neutral words by augmenting the corpus with some intervention mechanism.

For instance, to address the gender imbalance in training corpora like OntoNotes 5.0, Zhao et al. (2018a) propose a training-time data augmentation technique in which all male entities are substituted for female entities and vice versa. In this rule-based approach, all named entities are first anonymized, and then a set of rules (e.g., "she → he", "Mr." → "Mrs.", "mother" → "father") would be applied to all matching tokens in OntoNotes 5.0 to produce a new corpus. For instance, a sentence like "John went to his house" would become its counterfactual counterpart "E1 went to her house" after gender swapping. Coreference systems would then be trained on the union of the original corpus and the gender-swapped corpus. In the study (Zhao et al., 2018a), by adopting this gender-swapping data augmentation approach in combination with debiased word embeddings, the end-to-end neural coreference system and the feature-rich system both pass the WinoBias test.

In addition to the simple heuristic of swapping gendered word pairs, Lu et al. (2018) introduce more nuanced adjustments in order to maintain semantic and grammatical structures. For instance, instead of flipping all gendered words, the proposed model (called the *grammatical intervention*) uses coreference information to avoid modifying words that are in a coreference chain with a proper noun (e.g., Queen Elizabeth). This would prevent a sentence like "Elizabeth … she … queen" from being used to generate a counterfactual sentence like "Elizabeth … he … king" (Maudslay et al., 2019). Furthermore, the model uses part-of-speech tags (which are part of the training corpus metadata) to disambiguate between the objective pronoun "her" (which would map to "him") and the possessive pronoun "her" (which would map to "his").

While the paradigm proposed by Lu et al., (2018) can successfully mitigate direct bias, it is less effective when reducing indirect bias. In response to the approach of skipping words that refer to proper nouns, Maudslay et al. (2019) propose an alternative called the Names Intervention, which explicitly addresses first names and the inherent bias associated with them. For instance, by keeping sentences like "Tom … He is a successful and powerful executive" in the corpus with no counterfactual counterpart, the stereotypical association between "he" and "executive" would persist in the augmented corpus. Therefore, the Names Intervention uses bipartite matching to pair the 2500 most common names in the United States Social Security Administration (SSA) dataset[8] based on frequency and the degree of gender specificity. For instance, a common male name such as *John* may be paired with a common female name like *Mary*, while a more gender-neutral name like *Taylor* may be paired with a name like *Jordan* via the bipartite matching algorithm. With this variant of CDA, Maudslay et al. are able to mitigate not only direct but also indirect bias.

[8] https://www.ssa.gov/oact/babynames/background.html

While the CDA variants described above yield a more balanced corpus, they are only applicable to English and other languages with limited morphological inflection. For morphologically rich languages, simply swapping gendered words might lead to gender disagreement and produce ungrammatical sentences. To tackle this challenge, Zmigrod et al., (2019) propose using a Markov random field to infer a new morpho-syntactic tag sequence after an intervention on the grammatical gender of a word. By reinflecting entire sentences, this variation of CDA produces corpora with higher grammaticality than the naïve swapping approach.

3.3.3 Dropout Regularization

Dropout regularization (Srivastava et al., 2014) is a technique for reducing overfitting when training large neural models. By randomly ignoring some units during training, dropout regularization helps to prevent neurons from becoming overly reliant on the specialization of their neighboring neurons. While it is part of the training process for BERT, it is not applied in ALBERT since it has fewer parameters.

Using the two existing dropout parameters in BERT – one for activation weights, a, and another for hidden activations, h – Webster et al. (2020) experiment with increasing the dropout values for BERT from the default (i.e., $a = 0.1$ and $h = 0.1$) to $a = 0.15$ and $h = 0.20$. As a result, the gendered correlations between words are reduced correspondingly, with the Terms DisCo value decreased from 1.0 to 0.0 and the Names DisCo value increased from 3.4 to 0.7. As for ALBERT, introducing dropout values $a = 0.05$ and $h = 0.05$ (best results from grid search) leads to a substantial decrease in gendered correlations when evaluated on the Winogender (Rudinger et al., 2018) challenge set, even though the DisCo values do not reflect the same improvement. Nevertheless, Webster et al. conclude that it is valuable to incorporate dropout regularization in model configuration since it may help reduce unintended correlations that are not necessarily captured in accuracy metrics.

IV. Gender Bias in Neural Machine Translation

This section delves into the problem of gender bias in NMT. It first identifies some obstacles for machine translation, such as pronoun-dropping and genderless pronouns. It then examines the challenge dataset WinoMT, which has become an influential framework for evaluating gender bias in machine translation. Lastly, the section discusses various approaches for mitigating such bias.

4.1 Challenges in Machine Translation

Among the various aspects that make machine translation challenging, the phenomenon of pronoun-dropping is one such factor that contributes to the complexity of this task. In a pro-drop language like Japanese or Chinese (Wang et al., 2018), pronouns can be omitted when they are inferable from the context. Furthermore, the conditions in which this may occur are intricate, and the frequency of omission varies from language to language. Spanish, for example, also exhibits pro-drop properties, but to a lesser extent than Japanese and Chinese. Therefore, when translating from a pro-drop language to a non-pro-drop language like English, NMT models have to infer the invisible pronouns from contextual information. Since translation proceeds on a sentence-level, however, critical contextual cues might simply be unavailable at decoding time to generate an accurate translation.

To highlight the lack of document-level consistency, consider the following excerpt from Britney Spears' Spanish Wikipedia page[9] (Webster & Pitler, 2020): "Britney Jean Spears... **Ø** Adquirió fama durante **su** niñez al participar en el programa de televisión The Mickey Mouse Club (1992)." This sentence contains a dropped subject and a neutral possessive noun, all referring to "Britney Jean Spears", which appears earlier in the text. However, popular NMT systems like Google Translate are inclined to use masculine pronouns and produce a translation like "**He** gained fame during **his** childhood by participating in the television program The Mickey Mouse Club (1992)." Moreover, since inference is made in a left-to-right fashion, the effect of a mistranslation might cascade and contribute to further mistakes in the translation, undermining the overall BLEU score (Saunders & Byrne, 2020).

Another challenge in translation is the expression of gender. For instance, since languages like Spanish and French have a more elaborate grammatical gender system than English, these languages differ in their ability to encode gender information (Vanmassenhove et al., 2018). While a French speaker may utter either "Je suis heureux" or "Je suis heureuse" depending on their gender, the English equivalent would be "I am happy", which contains no gender information about the speaker. Therefore, to produce a grammatically correct translation from English to a language like French, a translator has to somehow recover the speaker's

[9] Source: https://es.wikipedia.org/wiki/Britney_Spears.Translation: Retrieved from translate.google.com, Feb 27, 2019 (Webster & Pitler, 2020).

gender. On the other end of the spectrum, there exist genderless languages, which do not distinguish between grammatical genders. While human translators can often obtain or infer the underlying gender from contextual information, most MT systems rely on sentence-level statistical dependencies that have been learned from training. As a result, they are less attuned to the broader context and often fail to correctly determine the gender trait of the original speaker. For instance, the Hungarian pronoun ő can refer to "he", "she", or "it" according to the context. A case study on Google Translate (Prates et al., 2019) demonstrates that the NMT model is prone to translating ő to "she" when it describes a nurse, and to "he" when it describes a CEO, as depicted in Figure 4.1.

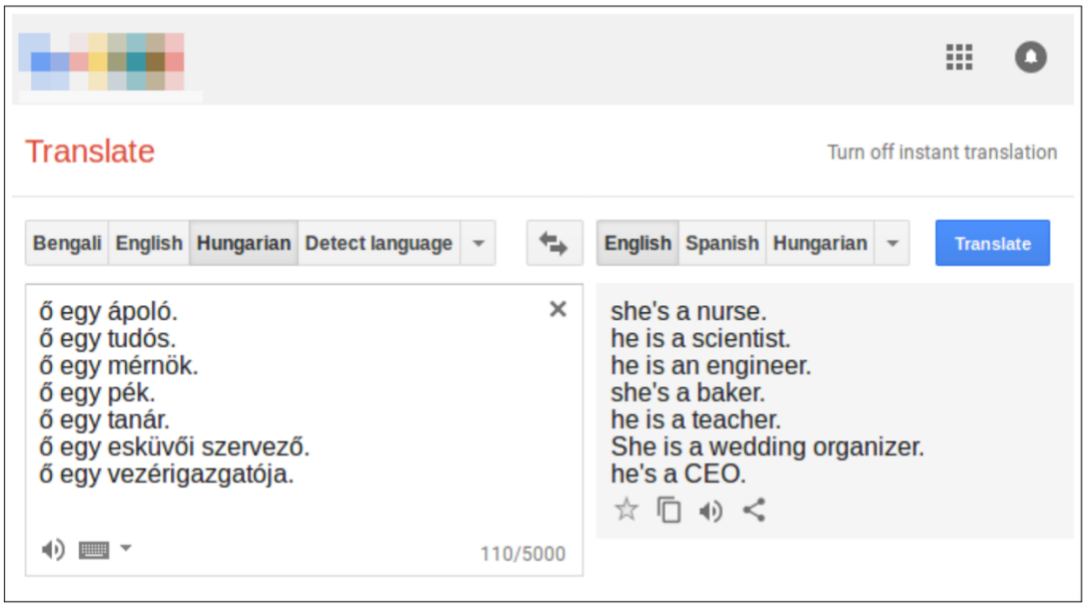

Figure 9 Translating Hungarian sentences with the genderless pronoun ő into English (Prates et al, 2019).

In addition, gender-neutral language has become more widely adopted in languages such as English. Unlike genderless language, which simply does not differentiate between natural genders, gender neutrality refers to word choices that do not presuppose a particular natural gender. The use of gender-neutral language, such as the singular *they*, is both an effort to promote inclusivity and avoid word choices that reflect stereotypes baked into gendered language; however, this also creates ambiguities that MT systems often grapple to resolve.

4.2 Evaluation Frameworks

4.2.1 WinoMT

To evaluate bias in state-of-the-art machine translation (MT) models, Stanovsky et al. (2019) introduce a large-scale multilingual challenge corpus called WinoMT[10]. WinoMT is built upon

[10] https://github.com/gabrielStanovsky/mt_gender

the WinoBias dataset (Zhao et al., 2018a) and the Winogender dataset (Rudinger et al., 2018). Similar to the coreference resolution task, an accurate translation also depends on correct gender identification. A system prone to gender bias might overlook contextual cues (e.g., morphological markers) and assign gender – or in the case of the coreference task, resolve pronouns – based on social stereotypes. For instance, given the Spanish source sentence *"La doctora le pidió a la enfermera que le ayudara con el procedimiento"*, in which the doctor is a woman based on the feminine article ("*la*") and the feminine inflection ("*-a*" in "*doctora*"), a biased MT system might mistranslate the doctor as male.

Given this parallel between coreference resolution and machine translation, WinoMT is constructed as a concatenation of WinoBias and Winogender, with a total of 3,888 sentences. The challenge set contains a similar number of sentences with the male and the female gender (1826 and 1822, respectively), as well as 240 gender-neutral sentences. Furthermore, the dataset is balanced between stereotypical and non-stereotypical gender-role assignments (e.g., a male doctor vs. a male nurse).

The evaluation framework consists of an automatic translation of WinoMT sentences to eight target languages, which belong to four different language families (i.e., Romance, Slavic, Semitic, Germanic) and embody a wide range of linguistic properties (in terms of word order, grammar etc.). To detect and quantify bias in an MT model, denoted *M*, all sentences in the challenge set are first translated into a target language, denoted *L*, via *M*, producing a bilingual corpus of English and *L*. An alignment technique called *fast_align*[11] (Dyer et al., 2013) is then applied to the bilingual corpus, mapping the annotated English entity (e.g., "the doctor") to its translation (e.g., "el doctor"). In the final step of the pipeline, the gender of each translated sentence is extracted via morphological analysis and compared against the gold annotation in the English dataset.

For the study, Stanovsky et al. use WinoMT to assess four widely used commercial MT systems – Google Translate[12], Microsoft Translator[13], Amazon Translate[14], SYSTRAN[15] – in addition to two state-of-the-art academic models, developed by Ott et al. (2018) and Edunov et al. (2018), respectively. In terms of accuracy, which is defined as the percentage of instances in which the translation correctly predicts the gender of the entity, most systems perform no better than a random guess, with accuracy around or less than 50%. Even though some systems perform better on English-to-German translation – with Microsoft Translator achieving 74.1% in accuracy, the authors suggest that this could be attributed to German's similarity to English.

In terms of the F_1 score, all systems – except Microsoft Translator on German – achieve a significantly higher score when translating sentences with male instances than female instances. Last but not least, all six systems exhibit inconsistency when evaluating sentences with stereotypical and non-stereotypical gender roles. Similar to the performance of Google Translate

[11] https://github.com/clab/fast_align
[12] https://translate.google.com/
[13] https://www.bing.com/translator
[14] https://aws.amazon.com/translate/
[15] https://www.systransoft.com/

(depicted in Figure 4.1), all systems achieve higher accuracies (and F_1 scores) when translating sentences with stereotypical gender role assignments (e.g., a female nurse) than non-stereotypical assignments (e.g., a male receptionist).

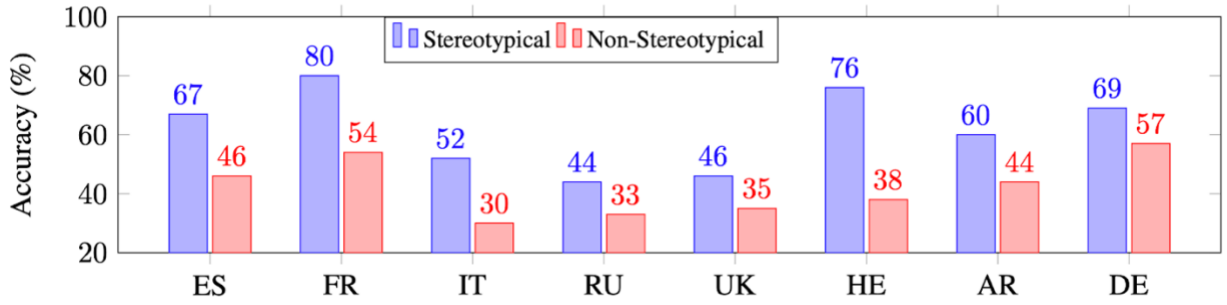

Figure 10 Performance of Google Translate when translating from English to eight target languages (Stanovsky et al., 2019).

As the first large-scale multilingual challenge set for NMT, the WinoMT metrics have established a baseline for subsequent studies, including the work by Saunders and Byrne (2020) on bias reduction. However, this evaluation framework has several limitations. As a combination of WinoBias and Winogender, WinoMT consists solely of artificially created sentences. While this creates a controlled experiment environment, it might introduce unintended biases that risk interfering with the evaluation process. Furthermore, WinoMT is limited both in size and scope. It neither reflects the normal size of a natural language processing task nor captures the range of domains machine translation is used in. As a result, the WinoMT challenge set only serves as a proxy for detecting and quantifying gender bias in machine translation.

4.3 Mitigation Techniques

To test how susceptible MT models are to gender bias, Stanovsky et al. (2019) experiment with a "fighting bias with bias" approach in which they prepend adjectives to occupation words in the WinoMT dataset. Adjectives like "handsome" and "pretty", which are often associated with gender-specific nouns, may be used to describe stereotypically female or male professions; for instance, "The doctor asked the nurse to help her in the operation" would be converted to "The *pretty* doctor asked the nurse to help her in the operation". While this approach leads to notable improvements in languages such as Spanish, Russian, and Ukrainian – with 11.2% increase in accuracy for Russian – Stanovsky et al. acknowledge that this technique is limited in scope. Not only is this debiasing scheme difficult to be generalized, but it also assumes an accurate coreference system that always correctly resolves pronouns. Nevertheless, this experiment serves to illustrate how MT systems can be easily influenced by the connotations of words and are prone to bias.

In addition to the experiment described above, this section explores methods that have been proposed to tackle gender bias in machine translation. While Section 4.3.1 focuses on

techniques derived from mitigation frameworks used on word embeddings and other pre-trained models (Section 3.3), Section 4.3.2 discusses a novel, and arguably more efficient, approach that involves fine-tuning as opposed to retraining MT models.

4.3.1 Word Embeddings Techniques

Inspired by the role of debiased word embeddings in reducing bias in many NLP tasks, Escudé Font and Costa-jussà (2019) experiment with a similar approach for NMT. Their methodology entails employing different variations of the Global Vectors (GloVe) embeddings for the encoder and decoder, and then evaluating each model's performance with the BLEU metric. For this experiment, the encoder and decoder in the OpenNMT Transformer[16] architecture are pre-trained with either the original GloVe, hard-debiased GloVe (via the debiasing process described in Section 3.3.1), or Gender-Neutral GloVe (GN-GloVe; Zhao et al., 2018b), while the Transformer model without any pre-trained embeddings serves as the baseline. Furthermore, for each type of embedding, three possible cases are tested: using embeddings only for the encoder, only for the decoder, and for both the encoder and decoder.

The training corpus consists of over 16 million pairs of English-Spanish sentences from an amalgam of sources, such as the United Nations[17], Europarl (Koehn, 2005), CommonCrawl[18], and the Workshop on Machine Translation (WMT)[19]. To assess the effectiveness of word embeddings, *newstest2013* – a test set of 3000 sentences provided by WMT – is utilized. To investigate gender bias in the resulting MT systems, Escudé Font and Costa-jussà develop an additional test set with sentences of the pattern "*I've known* {*her, him,* <proper noun>} *for a long time, my friend works as {a, an}* <occupation>." in which <proper noun> refers to a proper name such as "John" or "Mary", and <occupation> is selected from a list of professions provided by the U.S. Bureau of Labor Statistics.

When evaluated on the *newstest2013* test set, the model with GN-GloVe in both the encoder and decoder exceeds the baseline by 0.98 in BLEU score. Meanwhile, to analyze the extent of gender bias in these MT systems, Escudé Font and Costa-jussà focus on the translation of "friend" – which could be either "amiga" (feminine) or "amigo" (masculine) depending on the context. The authors observe that with "him" or "John", "amigo" is almost always predicted at 100% accuracy for all models. While most models are able to predict "amiga" based on "her" with high accuracy, many struggle when attempting to predict "amiga" based on "Mary". Among the various settings, hard-debiased GloVe – when applied to both the encoder and decoder – performs best across all scenarios. In addition to "friend", the translation of occupation words in the context of "her" also improves when GN-GloVe is used in both the encoder and decoder, especially for technical roles like "criminal investigator", "heating mechanic", and "refrigeration mechanic".

[16] https://opennmt.net/
[17] https://conferences.unite.un.org/uncorpus
[18] https://commoncrawl.org/
[19] http://www.statmt.org/wmt13/

However, similar to the WinoMT evaluation framework, a shortcoming of this study on gender bias is that it is limited to the domain of professional occupations. As a result, its findings may not be generalizable to other fields in which machine translation is applied. Moreover, the study is only concerned with English-Spanish translation, which does not capture the challenges of translating languages with different linguistic properties.

4.3.2 Domain Adaptation Techniques

While popular bias mitigation techniques often involve synthetic gender-balanced corpus or debiased word embeddings, Saunders and Byrne (2020) introduce an alternative approach that relies on fine-tuning rather than retraining. They argue that not only is it time-consuming to train an MT model from scratch, but it also presupposes that the source of bias in a corpus can be easily identified and remedied. In comparison, domain adaptation – which relies on a small portion of in-domain data to calibrate an NMT model – is much more efficient. By positioning bias mitigation as a domain adaptation problem, Saunders and Byrne seek to debias NMT models via a small, gender-balanced adaptation set as well as a counterfactual set during fine-tuning.

Similar to prior studies on NLP gender bias (Zhao et al., 2018a, Rudinger et al., 2018), the authors use a dataset of coreference sentences containing occupation words to reveal potential bias in a system. Based on 194 professions from the US labor statistics, 388 sentences are created based on the template "*The* [Profession] *finished* {*his, her*} *work*". These sentences constitute a handcrafted set that is then manually translated into three target languages, namely German, Spanish, and Hebrew, representing three distinct language families. Furthermore, Saunders and Byrne develop a set of counterfactual data that is augmented via gender-swapping, as described by Zhao et al. (2018a). The authors then use a general NMT model to forward-translate the gender-swapped English source sentences into corresponding gender-swapped target sentences.

To train the general-purpose models, the authors rely on three large corpora of bilingual data: WMT19 news task datasets (Barrault et al., 2019) for English-German, United Nations Parallel Corpus for English-Spanish, and multilingual TED talks corpus (Cettolo et al., 2014) for English-Hebrew. For all three sources, around 11-12% of the datasets are gendered sentences, with slightly more sentences involving male entities than female entities.

To assess the effectiveness of fine-tuning on the counterfactual data and the handcrafted profession data, Saunders and Byrne compare the outcomes against the baseline results from WinoMT study (Stanovsky et al., 2019). Adapting the model to a gender-swapped corpus not only increases the accuracy for English-German and English-Spanish, but it also decreases the difference in F_1 score between translation of male and female entities (denoted ΔG), as well as the difference in F_1 score between pro-stereotypical and anti-stereotypical gender role assignments (denoted ΔS), for these two language pairs. Even though this trend is promising, the improvement is rather subtle and is not replicated for English-Hebrew.

On the other hand, adapting NMT models to the set of sentences focused on occupations produces more desirable results. As the handcrafted set is constrained in scope and format, fine-tuning on this dataset only takes a few minutes on a single GPU, compared to several hours required for the counterfactual dataset. While the handcrafted set leads to an 19% increase in accuracy from the WinoMT baseline and more substantial improvement in ΔG and ΔS than the counterfactual set, it results in a drop in BLEU score.

This degradation of general translation quality can be attributed to catastrophic forgetting, which is the enhancement of in-domain knowledge at the expense of general domain knowledge. To overcome this pitfall, Saunders and Byrne (2020) propose using regularized training (Barone et al., 2017) and lattice rescoring to minimize the tradeoff. In particular, the lattice rescoring approach not only maintains (and sometimes even enhances) the general BLEU score, but it also facilitates debiasing and yields better performance on WinoMT metrics.

Based on these experiments, Saunders and Byrne conclude that domain adaptation may be an efficient and effective strategy for reducing gender bias in NMT. The study also illustrates that fine-tuning on a small, handcrafted gender-balanced dataset may be more fruitful than a counterfactual dataset. Moreover, procedures such as lattice rescoring can be employed to preserve the translation quality during fine-tuning, allowing NMT systems to be debiased without compromising its overall performance.

4.3.3 Cross-lingual Pivoting Technique

To address the challenges presented by implicit and ambiguous pronouns, Webster and Pitler (2020) propose using a cross-lingual pivoting technique to automatically produce gender labels that facilitate pronoun translation.

This approach is partially inspired by previous work on adding tags to explicitly convey the gender of first-person singular pronouns (Vanmassenhove et al., 2018). In particular, Vanmassenhove et al. are interested in the translation from English to 10 languages. The target languages include French, Portuguese, Italian, Spanish, and Greek – all of which require morphological agreement with the gender of the speaker. To train gender-informed NMT models, all English source sentences in the parallel corpora are enriched with tags expressing the gender of the speaker, as illustrated in "FEMALE Madam President, as a…" When evaluated on a male-only and a female-only test set containing first-person singular pronouns, the gender-informed NMT systems demonstrate the greatest improvement on the female set. In addition, among the five target languages with grammatical gender agreement, all (except for Spanish) experience an increase in BLEU score when evaluated even on a general test set, as indicated in Table 1. This suggests that gender-informed systems not only enhance morphological agreement, they also better capture other subtle ways in which gender identity might be manifested, such as sentence constructions and word preferences.

Systems	EN	EN-TAG
FR	37.82	**39.26***
ES	42.47	42.28
EL	31.38	**31.54**
IT	31.46	**31.75***
PT	36.11	**36.33**
DA	36.69	**37.00***
DE	28.28	28.05
FI	21.82	21.35*
SV	35.42	35.19
NL	28.35	28.22

Table 1 BLEU scores for the 10 baseline (untagged) NMT systems (i.e., EN), and BLEU scores for the 10 gender-informed NMT systems (i.e., EN-TAG). Statistically significant differences ($p < 0.05$) are marked by * (Webster & Pitler, 2020).

Building on this approach of incorporating gender information into training datasets, Webster and Pitler devise a method to automatically enrich corpora with gender tags. While this study primarily focuses on English-Spanish translation, the technique is language agnostic and can be extended to other language pairs with different linguistic properties.

Using large corpora of English and non-English (e.g., Spanish) Wikipedia pages, the multi-lingual pivot extraction pipeline consists of three stages:

1. **Page Alignment**: pairs of pages in English and Spanish with the exact same title are identified.
2. **Sentence Alignment**: pairs of sentences that express approximately the same meaning are identified. This process involves translating sentences in the Spanish page to English, and then performing bipartite matching over these English translations and sentences from the English page. In addition to being a one-to-one mapping, each pair must share either a noun or verb, and the edit sentence is at most one half of the sentence length.
3. **Pronoun Tagging**: for each sentence pair, perform alignment over the tokens to identify cases of dropped and gender-neutral pronouns (e.g., Spanish possessive pronoun *su*). Use the gender of the English pronoun (i.e., *she*. *her, he, his*) as a label for the ambiguous target pronoun.

For example, given the English[20] and Spanish Wikipedia[21] articles on "Mitsuko Shiga" (obtained via Page Alignment), the labeled sentence produced by Pronoun Tagging is "**Ø/She** Publicó numerosas antologas de **su/her** poesa durante **su/her** vida, incluyendo Fuji no Mi, Asa Tsuki, Asa Ginu, y Kamakura Zakki."

Finally, to account for the potential disparity in representation by gender in Wikipedia (Wagner et al., 2016), masculine examples are down-sampled to obtain a gender-balanced dataset, which consists of 79,240 prodrop and 187,224 possessive examples. This dataset is then employed to fine-tune BERT (Devlin et al., 2019), a state-of-the-art pretrained language model. To assess the effect of cross-lingual pivoting, the authors use two types of input – single

[20] https://en.wikipedia.org/wiki/Mitsuko_Shiga
[21] https://es.wikipedia.org/wiki/Mitsuko_Shiga

sentences (denoted Sentences) and full sentences of up to 128 tokens (denoted Contexts) – to compare the performance of a baseline NMT model, a contextualized NMT model (in which Contexts is employed), and an NMT model augmented with fine-tuned BERT.

As the baseline, a Transformer (Vaswani et al., 2017) trained on WMT'13[22] Spanish-to-English data achieves an F_1 score of only 31-51% for feminine pro-drop instances. By masking the position of each dropped subject position (Devlin et al., 2019), the authors show that a non-fine-tuned BERT model achieves better F_1 scores than the baseline when producing gender predictions, particularly for feminine examples. Nevertheless, it is still prone to misclassifications, including predicting a masculine tag for the Spanish sentence: "Ø Adquirió fama durante **su** niñez al participar en el programa de televisión The Mickey Mouse Club (1992)." about Britney Spears. After training BERT over the gender-tagged dataset (generated by cross-lingual pivoting), the fine-tuned model demonstrates an improvement of over 20% in all instances, with a new F_1 score of 92% for both masculine and feminine pronouns.

To enhance machine translation quality, this fine-tuned BERT gender classifier can be integrated into the standard NMT architecture. In this study, all Spanish sentences containing a dropped or gender-neutral pronoun are annotated with gender tags predicted via BERT. For example, the Spanish sentence about Britney Spears is now extended to "Ø Adquirió fama durante **su** niñez al participar en el programa de televisión The Mickey Mouse Club (1992). <c> <FEM>", wherein <c> serves as a separator. As depicted in Table 2, incorporating gender tags into an NMT system leads to an 8.8% improvement in F1 score for feminine pronouns, surpassing the overall performance of both the baseline model and the contextualized MT model.

Model	Translation Input	Classifier Input	BLEU	Masculine			Feminine		
				P	R	F1	P	R	F1
Baseline MT	Sentences	-	34.02	95.2	**97.1**	96.2	69.7	57.5	63.0
+ Gender Tags	Sentences	Contexts	**34.12**	96.8	96.2	**96.5**	**70.0**	73.7	**71.8**
Context MT	Contexts	-	33.99	**97.6**	94.7	96.1	63.2	**80.0**	70.6

Table 2 BLEU score and prediction accuracy by gender on the WMT'13 Spanish-to-English test set (Webster & Pitler, 2020).

[22] http://www.statmt.org/wmt13/translation-task.html

V. Conclusion

As artificial intelligence (AI) technology becomes increasingly integrated into various facets of society, the potential of algorithmic tools to perpetuate and amplify social bias has received rapidly growing scrutiny (Mehrabi et al., 2019). As one of the most prominent subfields of AI research, NLP also grapples with the issue of algorithmic bias. Many NLP tasks, including machine translation, have diverse use cases ranging from law, finance, and healthcare to consumer applications. However, like other AI systems trained on human-generated data, NLP and MT models are also susceptible to human-like stereotypes and biases.

This thesis homes in on the issue of gender bias in machine translation. In the first section, it not only provides an overview of technical concepts such as word embeddings and neural networks that have revolutionized the field of MT, but it also introduces evaluation metrics such as BLEU and METEOR that can offer insight into the overall translation quality. The paper then presents a literature survey of gender bias in NLP and explores studies on bias in tasks like coreference resolution. In the subsequent section of the paper, we investigate gender bias in the context of machine translation. In addition to delineating some unique challenges with translation, we explore recent efforts to recognize and reduce translation errors associated with gender stereotypes. To analyze potential gender bias, researchers not only rely on performance metrics like the BLEU score but have also devised evaluation frameworks such as the WinoMT challenge set (Stanovsky et al., 2019); in the meantime, mitigation techniques such as domain adaptation (Saunders & Byrne, 2020) and cross-lingual pivoting (Webster & Pitler, 2020) have shown noteworthy improvements and laid important foundation for future research in this direction.

Nevertheless, the study of gender bias in machine translation – and NLP at large – is still a relatively nascent field. The following subsection outlines some challenges and possible future directions for this line of research.

5.1 Future Work

First and foremost, the term “bias” lacks a coherent and consistent definition in NLP literature. By surveying 146 papers on bias in NLP, Blodgett et al. (2020) conclude that the concept of “bias” is underspecified, and many papers fail to elucidate why the behavior described as “bias” is harmful, to which groups of people, and in what concrete ways. As a result of this terminological imprecision, many studies that claim to analyze “bias” in NLP differ not only in their motivations but also their metrics for determining progress (Maudslay et al., 2019). As an area for future work, Blodgett et al. (2020) emphasize that it is critical for NLP researchers to clearly articulate their definition of “bias” and engage with relevant literature outside of NLP to illuminate the downstream impact of such system behaviors.

Secondly, as a general shortcoming of work in this field, most studies focus exclusively on binary genders. This is a layered problem not only due to the distinction between natural

gender and grammatical gender in linguistics, but also because the concept of gender itself is complex and cannot be captured in a one-to-one correspondence with linguistic gender (Cao & Daumé III, 2020). As suggested by Tomalin et al. (2021), current NLP research often conflates linguistic and sociological gender. For instance, most gender-balanced datasets in the field (including some we discuss in this thesis) are only balanced with respect to the male/female dichotomy. As non-binary gender identities become more widely recognized and embraced, it is important to develop gender-inclusive language systems that offer users the flexibility to self-identify their preferred pronouns. As groundwork in this direction, Sun et al., (2021) demonstrate that a Transformer model can be trained with auto-generated corpora to produce gender-neutral English sentences (e.g., with singular *they*) with less than 1% word error rate.

In addition to creating adaptable models that can support various expressions of gender in English, future work should also explore advancing gender-inclusive models in non-English languages. In languages with concord systems that assign gender to parts of speech beside the noun (e.g., verbs, adjectives, determiners), further research is needed to ensure morphosyntactic agreement during mitigation processes such as counterfactual data augmentation. Furthermore, future work in this direction should account for the disparate ways in which non-binary gender is encoded in a language, which is shaped both by social and cultural contexts as well as by the morphological complexity of a language. For instance, genderless languages and languages with grammatical genders differ in their ability to convey gender-neutrality, and languages like Spanish might require morphological or lexical innovations to expand the expressiveness of the language (Savoldi et al., 2021).

Finally, to reveal the presence of bias in MT (and other NLP) systems, many studies rely on synthetic challenge sets (e.g., WinoMT) that are limited in scope and diagnostic power. They neither prove the absence of gender bias in a model (Rudinger et al., 2018) nor embody the diverse range of real-world scenarios that could introduce bias into a system. To develop a more holistic understanding of the issue, future studies would need to investigate the interplay between multiple identity categories (e.g., gender, race, socioeconomic background) and adopt an interdisciplinary approach to address the problem of bias in machine translation.

References

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR, abs/1409.0473*.

Banerjee, S., & Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *IEEvaluation@ACL*.

Barone, A.V., Haddow, B., Germann, U., & Sennrich, R. (2017). Regularization techniques for fine-tuning in neural machine translation. *EMNLP*.

Barrault, L., Bojar, O., Costa-jussà, M., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., & Zampieri, M. (2019). Findings of the 2019 Conference on Machine Translation (WMT19). *WMT*.

Bergsma, S., & Lin, D. (2006). Bootstrapping Path-Based Pronoun Resolution. *ACL*.

Blodgett, S.L., Barocas, S., Daum'e, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *ACL*.

Bolukbasi, T., Chang, K., Zou, J.Y., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *NIPS*.

Brown, P.F., Pietra, S.D., Pietra, V.D., & Mercer, R. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Comput. Linguistics, 19*, 263-311.

Caliskan, A., Bryson, J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*, 183 - 186.

Cao, Y.T., & Daumé III, H. (2020). Toward Gender-Inclusive Coreference Resolution. *ACL*.

Castilho, S., Moorkens, J., Gaspari, F., Sennrich, R., Sosoni, V., Georgakopoulou, Y., Lohar, P., Way, A., Barone, A., & Gialama, M. (2017). A Comparative Quality Evaluation of PBSMT and NMT using Professional Translators.

Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., & Federico, M. (2015). Report on the 11 th IWSLT Evaluation Campaign , IWSLT 2014.

Chakraborty, T., Badie, G., & Rudder, B. (2016). Reducing gender bias in word embeddings.

Cho, K., Merrienboer, B.V., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. *ArXiv, abs/1406.1078*.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT*.

Dorr, B., Snover, M., & Madnani, N. (2010). Part 5: Machine Translation Evaluation.

Durrett, G., & Klein, D. (2013). Easy Victories and Uphill Battles in Coreference Resolution. *EMNLP*.

Dyer, C., Chahuneau, V., & Smith, N.A. (2013). A Simple, Fast, and Effective Reparameterization of IBM Model 2. *HLT-NAACL*.

Edunov, S., Ott, M., Auli, M., & Grangier, D. (2018). Understanding Back-Translation at Scale. *EMNLP*.

Elman, J. (1990). Finding Structure in Time. *Cogn. Sci., 14*, 179-211.

Fellbaum, C. (1998). A Semantic Network of English: The Mother of All WordNets. *Computers and the Humanities, 32*, 209-220.

Font, J.E., & Costa-jussà, M. (2019). Equalizing Gender Biases in Neural Machine Translation with Word Embeddings Techniques. *ArXiv, abs/1901.03116*.

Gonen, H., & Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. *NAACL*.

Greenwald, A., McGhee, D., & Schwartz, J.L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology, 74 6*, 1464-80.

Harris, Z. (1954). Distributional Structure. *WORD, 10*, 146-162.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation, 9*, 1735-1780.

Joos, M. (1950). Description of Language Design. *Journal of the Acoustical Society of America, 22*, 701-707.

Junczys-Dowmunt, M., Dwojak, T., & Hoang, H.T. (2016). Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions. ArXiv, abs/1610.01108.

Jurafsky, D., & Martin, J. H. (2020). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3rd ed. draft). Pearson. Retrieved from https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation.

Koehn, P. (2017). Neural Machine Translation. *ArXiv, abs/1709.07809*.

Koehn, P. (2020). *Neural Machine Translation*. Cambridge: Cambridge University Press. doi:10.1017/9781108608480

Kurita, K., Vyas, N., Pareek, A., Black, A., & Tsvetkov, Y. (2019). Measuring Bias in Contextualized Word Representations. *ArXiv, abs/1906.07337*.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *ArXiv, abs/1909.11942*.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition.

Lee, K., He, L., Lewis, M., & Zettlemoyer, L. (2017). End-to-end Neural Coreference Resolution. *ArXiv, abs/1707.07045*.

Lu, K., Mardziel, P., Wu, F., Amancharla, P., & Datta, A. (2018). Gender Bias in Neural Natural Language Processing. *Logic, Language, and Security*.

Maudslay, R.H., Gonen, H., Cotterell, R., & Teufel, S. (2019). It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution. *ArXiv, abs/1909.00871*.

May, C., Wang, A., Bordia, S., Bowman, S.R., & Rudinger, R. (2019). On Measuring Social Biases in Sentence Encoders. *ArXiv, abs/1903.10561*.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. *ArXiv, abs/1908.09635*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *ICLR*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. NIPS.

Nalisnick, E.T., Mitra, B., Craswell, N., & Caruana, R. (2016). Improving Document Ranking with Dual Word Embeddings. *Proceedings of the 25th International Conference Companion on World Wide Web*.

Ott, M., Edunov, S., Grangier, D., & Auli, M. (2018). Scaling Neural Machine Translation. *WMT*.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. *ACL*.

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *NAACL-HLT*.

Prates, M.O., Avelar, P.H., & Lamb, L. (2019). Assessing gender bias in machine translation: a case study with Google Translate. *Neural Computing and Applications, 32*, 6363-6381.

Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., & Manning, C.D. (2010). A Multi-Pass Sieve for Coreference Resolution. *EMNLP*.

Rahman, A., & Ng, V. (2012). Resolving Complex Cases of Definite Pronouns: The Winograd Schema Challenge. *EMNLP-CoNLL*.

Rudinger, R., Naradowsky, J., Leonard, B., & Durme, B.V. (2018). Gender Bias in Coreference Resolution. *NAACL-HLT*.

Saunders, D., & Byrne, B. (2020). Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem. *ArXiv, abs/2004.04498*.

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., & Turchi, M. (2021). Gender Bias in Machine Translation. *ArXiv, abs/2104.06001*.

Schwartz, L. (2018). Chapter 8. The history and promise of machine translation. In *Innovation and expansion in translation process research* (pp. 161-190).

Snover, M., Dorr, B., Schwartz, R., & Micciulla, L. (2006). A Study of Translation Edit Rate with Targeted Human Annotation.

Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res., 15*, 1929-1958.

Stanovsky, G., Smith, N.A., & Zettlemoyer, L. (2019). Evaluating Gender Bias in Machine Translation. *ArXiv, abs/1906.00591*.

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding-Royer, E., Chang, K., & Wang, W.Y. (2019). Mitigating Gender Bias in Natural Language Processing: Literature Review. *ArXiv, abs/1906.08976*.

Sun, T., Webster, K., Shah, A., Wang, W.Y., & Johnson, M. (2021). They, Them, Theirs: Rewriting with Gender-Neutral English. *ArXiv, abs/2102.06788*.

Sutskever, I., Vinyals, O., & Le, Q.V. (2014). Sequence to Sequence Learning with Neural Networks. *NIPS*.

Tomalin, M., Byrne, B., Concannon, S., Saunders, D., & Ullmann, S. (2021). The practical ethics of bias reduction in machine translation: why domain adaptation is better than data debiasing. *Ethics Inf Technol*.

Vanmassenhove, E., Hardmeier, C., & Way, A. (2018). Getting Gender Right in Neural Machine Translation. *ArXiv, abs/1909.05088*.

Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. *ArXiv, abs/1706.03762*.

Wagner, C., Graells-Garrido, E., & García, D. (2016). Women through the glass ceiling: gender asymmetries in Wikipedia. *EPJ Data Science, 5*, 1-24.

Wang, L., Tu, Z., Shi, S., Zhang, T., Graham, Y., & Liu, Q. (2018). Translating Pro-Drop Languages with Reconstruction Models. *ArXiv, abs/1801.03257*.

Webster, K., Recasens, M., Axelrod, V., & Baldridge, J. (2018). Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. *Transactions of the Association for Computational Linguistics, 6*, 605-617.

Webster, K., & Pitler, E. (2020). Scalable Cross Lingual Pivots to Model Pronoun Gender for Translation. *ArXiv, abs/2006.08881*.

Webster, K., Wang, X., Tenney, I., Beutel, A., Pitler, E., Pavlick, E., Chen, J., & Petrov, S. (2020). Measuring and Reducing Gendered Correlations in Pre-trained Models. *ArXiv, abs/2010.06032*.

Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., & Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *ArXiv, abs/1609.08144*.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. (2018a). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. *ArXiv, abs/1804.06876*.

Zhao, J., Zhou, Y., Li, Z., Wang, W., & Chang, K. (2018b). Learning Gender-Neutral Word Embeddings. *EMNLP*.

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., & Chang, K. (2019). Gender Bias in Contextualized Word Embeddings. *ArXiv, abs/1904.03310*.

Zmigrod, R., Mielke, S.J., Wallach, H., & Cotterell, R. (2019). Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. *ArXiv, abs/1906.04571*.